PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2000-310993**
(43)Date of publication of application : **07.11.2000**

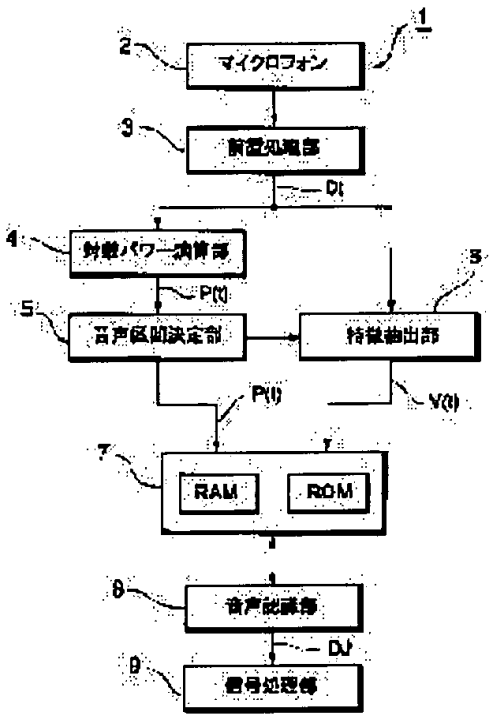
(51)Int.Cl. **G10L 11/02**

(21)Application number : **11-121457**
(22)Date of filing : **28.04.1999**

(71)Applicant : **PIONEER ELECTRONIC CORP**
(72)Inventor : **FUJITA IKUO**

(54) VOICE DETECTOR

(57)Abstract:
PROBLEM TO BE SOLVED: To recognize a voice by detecting and extracting a voice signal having no noise.
SOLUTION: Sound is collected by a microphone 2, a logarithmic power operation section 4 generates logarithmic power P(t) based on voice data Di. A voice section deciding section 5 compares levels of the logarithmic power P(t) based on a first threshold value of a higher level than a noise level of surrounding environment and a second threshold value of slightly higher than a noise level and a lower level than the first threshold value, the voice data when the logarithmic power P(t) varying temporally continuously to the higher level than the first threshold value is obtained out of the logarithmic power P(t) of the higher level than the second threshold value is detected as a uttered voice. And a feature extracting section 6 performs feature extracting based on the voice data detected as the uttered voice, makes a storage section 7 store the data of a feature vector V(t), further a recognizing section 8 recognizes a voice based on the data of the feature vector V(t), and the recognized result DJ is outputted to a signal processing section 9.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] Voice detection equipment characterized by providing the following. A sound detection means to change and output a sound to a sound signal A power conversion means to generate a power component signal of said sound signal The 1st threshold of predetermined level ** If level of said power component signal is compared based on the 2nd threshold of a low from said 1st threshold and a power component signal of a high level is detected from said 2nd threshold A voice section decision means to detect a sound signal which changes to a high level from the 1st threshold continuously in time among said sound signals as utterance voice, and to extract it, and a feature-extraction means to extract the feature information on audio based on said sound signal detected as utterance voice in said voice section decision means
[Claim 2] Said 1st threshold is voice detection equipment according to claim 1 which a high level and said 2nd threshold are higher than said noise level a little from noise level of perimeter environment, and is characterized by being set as a low from said 1st threshold.
[Claim 3] It is voice detection equipment according to claim 1 or 2 which is equipped with a storage means memorize said feature information which said feature-extraction means extracts, and is characterized by for said voice section decision section to make the feature information after a predetermined time before the time of the level of said power component signal exceeding said 1st threshold first among the feature information memorized by said storage means the feature information on said utterance voice.
[Claim 4] It has a storage means to memorize said feature information which said feature-extraction means extracts. Said voice section decision section Voice detection equipment according to claim 1 or 2 characterized by finally making the feature information before a future predetermined time into the feature information on said utterance voice from a time of level of said power component signal being less than said 1st threshold first among the feature information memorized by said storage means.
[Claim 5] It has a storage means to memorize said feature information which said feature-extraction means extracts. Said voice section decision section It is the feature information after a predetermined time before a time of level of said power component signal exceeding said 1st threshold first among said feature information memorized by said storage means. And voice detection equipment according to claim 1 or 2 characterized by finally making the feature information before a future predetermined time into the feature information on said utterance voice from a time of level of said power component signal being less than said 1st threshold first.
[Claim 6] Voice detection equipment given in any 1 term of claims 1-5 characterized by having a speech recognition means to perform speech recognition based on the feature information on said utterance voice.
[Claim 7] It is voice detection equipment according to claim 6 which said voice section decision means generates said power component signal synchronizing with a predetermined period decided beforehand, and is characterized by said speech recognition means performing said speech recognition synchronizing

with said voice section decision means setting up the feature information on said utterance voice.

[Translation done.]

* NOTICES *

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] This invention detects the uttered voice and relates to the voice detection equipment which removes and extracts a noise.
[0002]
[Description of the Prior Art] The integrated circuit device which makes high-speed signal processing possible is developed in recent years, and the voice recognition system using this integrated circuit device is being applied to electronic equipment. With the common voice recognition system, the feature extraction of the feature of the uttered voice is carried out, and the so-called speech recognition is performed by recognizing a vocabulary based on the extracted feature information. Although various voice-recognition algorithm is indicated in various kinds of reference in order to raise the rate of speech recognition, it is necessary to detect the uttered voice faithfully as a premise for applying these algorithms, and to remove and extract a noise here. Moreover, to realize the man-machine system which enables the response of real time to a speaker, it is necessary to detect and extract voice at a high speed.
[0003] Drawing 8 is the mimetic diagram having shown the conventional voice detection method. a logarithm [in / by collecting the uttered voice (only henceforth voice) with a microphone in this drawing (a), and searching for the logarithm of the addition value, while integrating the electrical signal (sound signal) acquired by this for every predetermined period / every above-mentioned predetermined period of a sound signal] -- power P (t) is generated.
[0004] Change of power P (t) is compared with the predetermined threshold THD. and the logarithm generated for every above-mentioned predetermined period -- a logarithm -- by judging the section which serves as a low from the voice section and a threshold THD in the section when the level of power P (t) becomes larger than a threshold THD to be the noise section, a voice component and a noise component are distinguished and it is extracting as a voice component which had only the sound signal within the voice section uttered.
[0005] namely, the logarithm among the sound signals acquired by sound-collecting -- the sound signal acquired in the section when the level of power P (t) became higher than a threshold THD -- as a true voice component -- distinguishing -- a logarithm -- he was trying to remove it noting that the sound signal acquired in the section when the level of power P (t) became lower than a threshold THD was a noise component
[0006]
[Problem(s) to be Solved by the Invention] However, the above-mentioned conventional voice detection method was not enough as the distinction precision of a voice component and a noise component. For this reason, it was difficult to detect a voice component faithfully and to extract it.
[0007] For example, although it will contribute to avoiding lack of a voice component when a threshold THD is set up low and the gap of te is extended at the ts and termination time at the initiation time of the voice section, as shown in drawing 8 (b) the logarithm which came out on the other hand and included

the noise -- since the criterion of power P (t) would fall, there was a problem of carrying out the misjudgment law of the noise in a sound signal to a voice component, and extracting it.
[0008] On the other hand, since the voice section (audio logging section) would narrow if a threshold THD is made high in order to avoid extracting the noise in a sound signal accidentally as shown in drawing 8 (c), there was a problem of it becoming impossible to extract the required voice component in a sound signal faithfully.
[0009] This invention is made in order to conquer the above-mentioned conventional trouble, and while removing a noise and detecting and extracting a voice component with a sufficient precision, it aims at offering the voice detection equipment which performs detection and an extract at a high speed.
[0010]
[Means for Solving the Problem] A sound detection means for this invention to change a sound into a sound signal, and to output in order to attain the above-mentioned purpose, A power conversion means to generate a power component signal of the above-mentioned sound signal, and the 1st threshold of predetermined level **, If level of the above-mentioned power component signal is compared based on the 2nd threshold of a low from the 1st threshold of the above and a power component signal of a high level is detected from the 2nd threshold of the above A voice section decision means to detect a sound signal which changes to a high level from the 1st threshold continuously in time among the above-mentioned sound signals as utterance voice, and to extract it, It considered as a configuration possessing a feature-extraction means to extract the feature information on audio based on the above-mentioned sound signal detected as utterance voice in the above-mentioned voice section decision means.
[0011] When according to this configuration a sound signal which became the origin for generating the power component signal when a power component signal was set to a low from the 2nd threshold is removed as a noise and a power component signal is set to a high level from the 2nd threshold from a high level or the 1st threshold, it detects as a voice component which had a sound signal which becomes the origin for generating the power component signal uttered, and extracts.
[0012] Here, when it is again set to a low from the 2nd threshold after a power component signal was set to a high level from the 2nd threshold, a sound signal at that time is removed as a noise. That is, when a power component signal used as a high level is continuously set to a high level from the 2nd threshold from the 1st threshold in time, a sound signal of the continuous time amount within the limits is detected as original utterance voice, and is extracted. And based on a sound signal detected and extracted as original utterance voice, a feature-extraction means extracts the feature information on audio.
[0013] Moreover, it had a storage means memorize the above-mentioned feature information which the above-mentioned feature-extraction means extracts, and the above-mentioned voice section decision section considered the feature information after a predetermined time before the time of the level of the above-mentioned power component signal exceeding the 1st threshold of the above first among the feature information memorized by the above-mentioned storage means as the configuration which sets up as feature information on the above-mentioned utterance voice.
[0014] According to this configuration, a case where it is continuously set to a high level from the 1st threshold in time while a power component signal with which a noise of a high level is included from the 2nd threshold, and the noise was included in a power component signal set to a high level from the 2nd threshold has not been set to a low from the 2nd threshold is detected. And more finally than it, a former sound signal after a predetermined time is made into an utterance signal on the basis of a time of a power component signal serving as a high level from the 1st threshold at the beginning. Consequently, even when a noise mixes in first transition of utterance voice, mixing of a noise can be suppressed to the minimum and original utterance voice can be extracted.
[0015] Moreover, it had a storage means memorize the above-mentioned feature information which the above-mentioned feature-extraction means extracts, and the above-mentioned voice section decision section considered the feature information before a future predetermined time as the configuration which finally sets up as feature information on the above-mentioned utterance voice from the time of the level of

the above-mentioned power component signal being less than the 1st threshold of the above first among the feature information memorized by the above-mentioned storage means.
[0016] According to this configuration, when a power component signal set to a high level from the 1st threshold is again set to a low (however, the 2nd threshold high level) from the 1st threshold, a case where a noise mixes in a sound signal at that time is detected. And on the basis of a time of a power component signal serving as a low from the 1st threshold at the beginning, from it, a subsequent predetermined time is set up and, finally a sound signal before the time is made into an utterance signal. Consequently, even when a noise mixes in a trailing edge of utterance voice, mixing of a noise is suppressed to the minimum and original utterance voice is extracted.
[0017] It has a storage means to memorize the above-mentioned feature information which the above-mentioned feature-extraction means extracts. Moreover, the above-mentioned voice section decision section It is the feature information after a predetermined time before a time of level of the above-mentioned power component signal exceeding the 1st threshold of the above first among the above-mentioned feature information memorized by the above-mentioned storage means. And the feature information before a future predetermined time was considered as a configuration finally set up as feature information on the above-mentioned utterance voice from a time of level of the above-mentioned power component signal being less than the 1st threshold of the above first.
[0018] According to this configuration, a case where a noise mixes in first transition and a trailing edge of a sound signal by utterance of a speaker is detected, mixing of a noise is suppressed to the minimum, and a sound signal is extracted.
[0019] Moreover, it considered as a configuration equipped with a speech recognition means to perform speech recognition based on the feature information on the above-mentioned utterance voice. Moreover, the above-mentioned voice section decision means generated the above-mentioned power component signal synchronizing with a predetermined period decided beforehand, and the above-mentioned speech recognition means was considered as a configuration which performs the above-mentioned speech recognition synchronizing with the above-mentioned voice section decision means setting up the feature information on the above-mentioned utterance voice. According to these configurations, speech recognition is performed whenever it extracts the feature information on utterance voice. By this, high-speed speech recognition is made possible.
[0020]
[Embodiment of the Invention] Hereafter, the gestalt of operation of this invention is explained with reference to drawing 1 thru/or drawing 8 . In addition, the voice detection equipment which enables voice actuation in the navigation system for mount using an audio equipment and GPS (Global Positioning System) navigation as 1 operation gestalt etc. is explained.
[0021] (Gestalt of the 1st operation) Drawing 1 is the block diagram showing the configuration of the voice detection equipment 1 of this operation gestalt. this voice detection equipment 1 -- the microphone 2 for sound-collecting, the front-end processing section 3, and a logarithm -- it has the power operation part 4, the voice section decision section 5, the feature-extraction section 6, the storage section 7, and the speech recognition section 8, and is constituted. The speech recognition section 8 outputs a recognition result to the signal-processing section 9 for operating the above-mentioned audio equipment etc. in addition, a logarithm -- the power operation part 4, the voice section decision section 5, the feature-extraction section 6, and the speech recognition section 8 are formed of the DIJITARI signal processor (Digital signal Processor:DSP) which operates according to the system program set up beforehand.
[0022] The pre amplifier which the front-end processing section 3 amplifies the electrical signal (raw sound signal) outputted from a microphone 2 here on the level in which signal processing is possible, and is outputted, The band pass filter which passes the frequency component in voice grade (for example, the range of 50Hz - 4kHz) among the above-mentioned sound signals outputted from pre amplifier, the A/D converter which changes into the digital voice data Di the sound signal which passed the band pass filter synchronizing with sampling frequency f more than a Nyquist rate (for example, f**11.025kHz) has --

having -- this voice data Di -- a logarithm -- the power operation part 4 is supplied.
[0023] a logarithm -- the voice data Di which produces the power operation part 4 synchronizing with sampling period deltaT (=1/f) -- predetermined every period Ts (for example, 10msec) -- integrating -- the addition value -- a logarithm -- calculating -- the logarithm for every predetermined period Ts of the above-mentioned sound signal -- power P (t) is generated as a power component signal, and is outputted. the logarithm of the square aggregate value of the voice data Di of the Tsxf individual (integer individual) obtained within each frame period Ts as it supposes that this period Ts is called a frame period and shown in a degree (1) -- a logarithm [in / for a value / each frame period Ts] -- it is referred to as power P (t).
[0024]
[Equation 1]

$$P(t)=\log\{\sum_{1}^{Ts\cdot f}(Di)^{2}\} \quad \cdots\cdots (1)$$

[0025] in addition, a logarithm -- the sign t of power P (t) is the coefficient of the integer which shows the sequence 1 and 2 in the time amount progress direction of each frame period Ts, 3 --, etc.
[0026] the voice section decision section 5 -- a logarithm -- the logarithm of the noise component contained in Power P (t) -- the logarithm of power and a voice component -- power (noise power and the power of a voice component are hereafter called speech power for the power of a noise component) is distinguished based on the thresholds THD1 and THD2 of two pieces. And the frame period Ts which speech power produces, and the frame period Ts which noise power produces are told by supplying the distinction result to the feature-extraction section 6. Moreover, the predetermined storage region of the storage section 7 mentioned later is made to memorize the data of number power of each sets P (t) generated in each frame period Ts.
[0027] The feature-extraction section 6 extracts the voice data Di of the Tsxf individual which exists within the frame period Ts which speech power produces based on the above-mentioned judgment result from the voice section decision section 5 (logging), and performs the feature extraction of a voice component by carrying out signal processing of such voice data Di. in addition, the LPC cepstrum (Linear Predictive coding Cepstrum) which is one of the linear predictive coding with this operation gestalt -- the feature extraction is performed based on law. That is, voice data Di is introduced into the voice generation model beforehand set up with the digital filter which has the linearity coefficient of about 20-dimensional one, and the feature extraction of the about 20-dimensional vector component predicted by the linear combination is carried out as feature data [of the spectral envelope of a voice component] (henceforth feature vector) V (t). And this feature-vector V (t) is supplied to the storage section 7 synchronizing with each frame period Ts.
[0028] Moreover, while suspending the above-mentioned logging processing about the voice data Di of the Tsxf individual which exists within the frame period Ts which noise power produces, processing of the above-mentioned feature extraction is also suspended. Therefore, feature-vector V (t) is outputted during the nascent state of a noise component.
[0029] that is, the voice section decision section 5 -- a logarithm, if the generating section of a voice component is distinguished based on power P (t) the distinction result -- being based -- the feature-extraction section 6 -- feature-vector V (t) -- generating -- outputting -- the voice section decision section 5 -- a logarithm, if the generating section of a noise component is distinguished based on power P (t) Since the feature-extraction section 6 does not generate feature-vector V (t) based on the distinction result, feature-vector [of a voice component] V (t) is supplied to the storage section 7, and a noise component is supplied to it.
[0030] The storage section 7 is equipped with the random access memory (RAM) which can restore, and the read-only memory (ROM) collating data was beforehand remembered to be, and is constituted.
[0031] the logarithm of the above [**** / storing various data temporarily in the case of the storage region MEM and speech recognition processing in which feature-vector V (t) transmitted from the feature-extraction section 6 synchronizing with a frame period Ts is memorized in order in the

above-mentioned RAM] -- the working area for memorizing the data of power P (t) etc. is assigned.
[0032] Two or more lexical information for collating a recognition result with the above-mentioned ROM is beforehand memorized as collating data. For example, various kinds of lexical information, such as "switch-on" for a speaker to perform voice actuation, "switch-off", "playback initiation", and "a halt", is memorized. Moreover, when this voice storage 1 is applied to the NABIGESHIN system for mount, the lexical information about geography, such as the name of a place and a name of the station, is also memorized.
[0033] The speech recognition section 8 recognizes the speech information of the uttered voice by collating feature-vector V (t) memorized by ****** MEM in RAM, and the collating data in ROM. And the data DJ of the recognition result is outputted to the signal-processing section 9.
[0034] Next, it explains with reference to the memory map shown in the wave form chart showing detailed actuation of the voice detection equipment 1 which has this configuration in the flow chart shown in drawing 2 , and drawing 4 , and drawing 5 . in addition, drawing 4 -- a logarithm -- change of power P (t) is shown typically and drawing 5 shows the memory map of a storage region MEM.
[0035] In drawing 2 , if voice detection equipment 1 starts, a microphone 2 will start sound-collecting irrespective of the existence of the utterance by the speaker (step S100). and the sound signal from a microphone 2 -- the front-end processing section 3 -- voice data Di -- changing -- further -- a logarithm -- the power operation part 4 -- every frame period Ts -- a logarithm -- power P (t) is generated and the voice section decision section 5 is supplied.
[0036] until directions of speech recognition initiation according [the voice section decision section 5 / on steps S102 and S104 and] to a speaker are made -- the logarithm for every frame period Ts -- power P (t) is measured in detail as noise level of perimeter environment. and -- criteria [P / (t) / number power of each sets] -- carrying out -- level slightly higher than it -- the 2nd threshold THD2 -- while making predetermined level higher than it into the 1st threshold THD1 on the basis of a threshold THD2 further -- a new logarithm -- whenever power P (t) is supplied, the 1st and 2nd threshold THD1 and THD2 is updated.
[0037] in addition, a logarithm -- the set point of the 1st and 2nd threshold THD1 and THD2 over power P (t) is beforehand decided experimentally in consideration of the electrical property of a microphone 2 or the front-end processing section 3. as an example -- a threshold THD2 -- a logarithm -- power P (t) -- about 5dB -- high -- a threshold THD1 -- a logarithm -- it is supposed that it will set up more highly about 10dB than power P (t).
[0038] If directions of speech recognition initiation are made (step S104), the 1st and 2nd threshold THD1 and THD2 for which the newest was asked will be decided, and speech recognition processing will be started (step S106).
[0039] next, the logarithm which the voice section decision section 5 inputs into every sampling period deltaT in step S108 -- the coefficient (positive integer) k for specifying the coefficient (positive integer) t for specifying the sequence of power P (t) and the address of a storage region MEM is set as t= 1 and k= 1. Thereby, the start address of a storage region MEM is specified.
[0040] next, the voice section decision section 5 -- a logarithm -- the logarithm from the power operation part 4 -- power P (t) -- inputting (step S110) -- a logarithm -- the value and the 2nd threshold THD2 of power P (t) are compared (S112).
[0041] here -- P (t) -- < -- the step S110 after counting up a coefficient t one time in THD2 (in the case of "YES") (step S113) -- returning -- the following logarithm -- power P (t) is inputted. On the other hand, in P(t) >=THD2 (in the case of "NO"), it shifts to step S114. namely, the logarithm produced by steps S110-S112 when the speaker has not yet spoken -- power P (t) is excepted from the processing object.
[0042] next, the logarithm which carried out [above-mentioned] the input in step S114 -- power P (t) is compared with the 1st threshold THD1. here -- the time of P(t) <THD1 (at the time of "NO") -- step S116 -- shifting -- a logarithm -- power P (t) is compared with the 2nd threshold THD2.
[0043] In P(t) >=THD2 (in the case of "YES"), in step S116, it shifts at step S117. The

feature-extraction section 6 calculates feature-vector V (t) based on the voice data Di within the P(t) >= frame period Ts applicable to the conditions of THD2, and makes a storage region MEM (k) memorize feature-vector V (t) here (step S118). namely, the logarithm which exceeded the 2nd threshold THD2 first -- feature-vector [in the frame period Ts applicable to power P (t)] V (1) is memorized in the storage region MEM of a start address (1).

[0044] next, the coefficients t and k -- respectively -- 1 -- counting up (step S120) -- the following logarithm -- after inputting power P (t) (step S122), processing of steps S114-S122 is repeated. Thereby, feature-vector V (t) which can be found for every frame period Ts is memorized in a storage region MEM (k).

[0045] however, repeat processing of these steps S114-S122 -- on the way -- alike -- step S122 -- setting -- the 2nd threshold THD2 -- the logarithm of a low -- when power P (t) is inputted, in step S116, it will judge with P(t) <THD2. That is, while a noise component will be judged, shifting to step S124 through the judgment "NO" of step S116 and counting up a coefficient t one time, after resetting a coefficient k to 1, the processing from step S110 is resumed substantially.

[0046] thus -- if steps S108-S124 are processed -- the 2nd threshold THD2 -- the logarithm of a high level, even if it is the case where power P (t) is inputted the 1st threshold THD1 -- the logarithm of a high level -- before inputting power P (t) -- the 2nd threshold THD2 -- the logarithm of a low, when power P (t) is inputted again As period tauns in drawing 4 shows, it judges that it is generated by all feature-vectors [in a storage region MEM (k)] V (t) based on a noise, and these feature-vectors V (t) is eliminated altogether. Consequently, a noise is appropriately removable.

[0047] Moreover, the ** a noise is not judged in step S116 to be while repeating processing of the above-mentioned steps S114-S122, the 1st threshold THD1 -- the logarithm of a high level, when power (t) is inputted (it is called the 1st case) The ** by which the judgment with a noise is not made in step S116 after the restart of the substantial processing from the above-mentioned step S110, the 1st threshold THD1 -- the logarithm of a high level -- the case (it is called the 2nd case) where power P (t) is inputted -- step S114 -- setting -- this logarithm -- power P (t) will be judged as P(t) >=THD1 ("YES" and judgment). And if it judges with "YES", it will shift to processing of step S126.

[0048] Thus, when it will shift to processing of step S126 with the 1st and 2nd above-mentioned case, feature-vector [from the initiation point in time A of the "voice logging section" in drawing 4 to the intermediate point in time B] V (t) will be memorized in an order from the start address of a storage region MEM (k) by the processing till then.

[0049] Furthermore, at the initiation time, since A becomes when a speaker speaks, it can be extracted without being missing in a voice component, and can be memorized to a storage region MEM (k). Furthermore, these-memorized all feature-vectors V (t) is higher than the average noise level in perimeter environment, and since [which was shown in period tauns] the noise of a high level is not included comparatively, either, it becomes data which does not include a noise.

[0050] in addition, the logarithm judged that is higher than the 1st threshold THD1 first in step S114 -- when power P (t) is the n-th thing, it is shown in the memory map of drawing 5 -- as -- the n-th - from the 1st -- feature-vector [of the 1st address] V (1) - V (n-1) become data with which are satisfied of the conditions of THD2 <=P(t) <THD1.

[0051] next -- if it shifts to step S126 -- a logarithm -- power P (t) is compared with the 2nd threshold THD2. Here, in P(t) >=THD2 (in the case of "YES"), it shifts to step S127 and the feature-extraction section 6 calculates the n-th feature-vector V (t) based on the voice data Di within the P(t) >= frame period Ts applicable to the conditions of THD2. Next, it shifts to step S128, and the above-mentioned feature-vector [of eye t=n watch] V (n) is memorized to the storage region MEM of the address of eye k=n watch (n), as shown in drawing 5 .

[0052] next, the step S130 -- setting -- coefficients t and k -- respectively -- 1 -- counting up -- further -- step S132 -- setting -- the following logarithm -- after inputting power P (t), it returns to step S126 and processing of steps S126-S132 is repeated.

[0053] Thus, if processing of steps S126-S132 is repeated, feature-vector V (n) called for within the

period to D from the time B of being shown in drawing 4 at the termination time - V (n+N) will be memorized in order by the storage region MEM (n) shown in drawing 5 - MEM (n+N). And feature-vector V (n+N) of a storage region MEM (n+N) becomes data when finally being judged with P(t) >=THD2 in step S126. Therefore, a noise component will be contained in feature-vector V (1) - V (n+N).

[0054] Next, after the speech recognition section 8 decides the time interval of the voice logging section in step S134 based on the total of feature data V (1) - V (n+N) and sampling period deltaT which were memorized to a storage region MEM (1) - MEM (n+N), In steps S136 and S138, the semantics of the language which the speaker uttered is recognized by collating feature-vector V (1) - V (n+N), and the collating data in ROM.

[0055] Next, the data DJ of the recognition result is outputted to the signal-processing section 9, and speech recognition processing is ended. In addition, after ending speech recognition processing of 1, it returns to step S100 again, and the same processing as the above is repeated.

[0056] the 2nd threshold THD2 which was set as a high level a little from the average noise level of perimeter environment according to this operation gestalt as stated above, and the 2nd threshold THD2 -- criteria [threshold / of high level / THD1 / 1st] -- carrying out -- a logarithm -- since the noise in Power P (t) and the original voice component were distinguished, a voice component can be extracted with high degree of accuracy.

[0057] Moreover, since feature-vector V (t) is extracted for every short-time frame period Ts, it becomes possible to carry out speech recognition on real time, and application to a man-machine system is possible.

[0058] (Gestalt of the 2nd operation) Next, the 2nd operation gestalt is explained with reference to the flow chart shown in drawing 6 . In addition, since the voice detection equipment of this operation gestalt is the same as that of the configuration shown in drawing 1 , the explanation about a configuration is omitted. Moreover, in drawing 6 , the sign same about a corresponding step of operation identically to drawing 2 is attached and shown.

[0059] When the difference between this operation gestalt and the 1st operation gestalt is described, this operation gestalt is in the point of having prepared step S129a and S129b among steps S128 and S130 while preparing step S119a and S119b among steps S118 and S120 in drawing 6 .

[0060] First, as the 1st operation gestalt explained, it sets to step S118. If feature-vector V (t) is memorized to a storage region MEM (k) next, it will set to step S119a and S19b. All feature-vectors V (t) memorized in the storage region MEM (k) until now and the collating data in ROM are collated, speech recognition is performed, and the data DJ of the recognition result is outputted to the signal-processing section 9. then, the coefficients t and k -- respectively -- 1 -- counting up (step S120) -- the further following logarithm -- after inputting power P (t) (step S122), it shifts to step S114.

[0061] Furthermore, as the 1st operation gestalt explained, it sets to step S128. If feature-vector V (t) is memorized to a storage region MEM (k) next, it will set to step S129a and S129b. All feature-vectors V (t) memorized in the storage region MEM (k) until now and the collating data in ROM are collated, speech recognition is performed, and the data DJ of the recognition result is outputted to the signal-processing section 9. then, the coefficients t and k -- respectively -- 1 -- counting up (step S130) -- the further following logarithm -- after inputting power P (t) (step S132), it shifts to step S126.

[0062] Thus, speech recognition is completed when the termination of a voice component is detected in step S126, since collating with collating data performs speech recognition whenever it memorizes a feature vector (t) to a storage region MEM (k).

[0063] For example, when a speaker utters saying "I want to go to the Meguro station", it sets to step S119a, S119b, and S119a and S119b. "**", "**", and "**" -- "-- obtaining -- " -- "-- coming -- " -- "-- passing -- " -- "-- it is -- " -- "-- coming -- " -- "-- **" -- "-- it is -- " -- ** -- when the word to say will be recognized in order and processing was "ended", the semantics of the language which the speaker uttered is decided. Consequently, according to this operation gestalt, detection, an extract, and speech recognition of a voice component can be extremely performed at a high speed.

[0064] (Gestalt of the 3rd operation) Next, the 3rd operation gestalt is explained with reference to the wave form chart shown in the flow chart shown in drawing 7 , and drawing 8 . In addition, since the voice detection equipment of this operation gestalt is the same as that of the configuration shown in drawing 1 , the explanation about a configuration is omitted. Moreover, in drawing 7 , the sign same about a corresponding step of operation identically to drawing 2 is attached and shown. Furthermore, drawing 8 is the wave form chart which matched with drawing 4 and was shown.
[0065] When the difference between this operation gestalt and the 1st operation gestalt is described, this operation gestalt is in the point of having prepared step S133a and S133b into the path which returns from step S132 to step S126 while forming step S125 among steps S114 and S126 in drawing 7 . Furthermore, it is characterized by the point of having prepared step S135a and S135b among steps S126 and S136.
[0066] first, the 1st operation gestalt explained -- as -- step S114 -- setting -- the 1st threshold THD1 -- the logarithm of a high level -- if power P (t) is judged -- step S125 -- setting -- this logarithm -- after memorizing the data of power P (t) to the working area in RAM by using that sequence t as the flag data FORWARD, it shifts to processing of step S126. therefore, the logarithm detected by B the time of being shown in drawing 8 -- the sequence t will be remembered to be power P (t) as flag data FORWARD.
[0067] furthermore, the logarithm which the voice section decision section 5 made the working area in RAM memorize in t-1 in step S133a in drawing 4 at the time before one from this time t -- Power P (t-1) -- reading appearance -- carrying out -- the logarithm -- Power P (t-1) is compared with the 1st threshold THD1. Here, after making the working area in RAM memorize in step S133b by using as the flag data BACKWARD sequence t-1 which is equivalent to t-1 at the time in P(t-1) >=THD1 (in the case of "YES"), it shifts to processing of step S126. On the other hand, it shifts to processing of step S126, without setting up the flag data BACKWARD in P(t-1) <THD1 (in the case of "NO").
[0068] the time of being shown in drawing 4 , if this processing was performed -- C, i.e., a logarithm, -- a logarithm [in / the time before / when power P (t) is again set to a low from the 1st threshold THD1 / one / C] -- the sequence t-1 of Power P (t-1) will be memorized as flag data BACKWARD.
[0069] And if processing moves from step S126 to step S135a, in step S135a, predetermined value (positive integer value) taua will be subtracted from the flag data FORWARD, and the subtraction result (=FORWARD-taua) will be set to ta at the presumed initiation time of a voice component. Furthermore, predetermined value (positive integer value) taud is subtracted from the flag data BACKWARD, and the subtraction result (=BACKWARD-taud) is set to td at the presumed termination time of a voice component.
[0070] here -- predetermined value taua and taud -- both -- a threshold THD2 -- the logarithm of a high level -- it has decided for power P (t).
[0071] Next, in step S135b, as shown in drawing 8 , the period between td(s) is decided with the last logging section Tc from ta at the presumed termination time at the presumed initiation time. And in step S136, after reading feature-vector V (t) which corresponds within the last logging section Tc from a storage region MEM (k) and collating with collating data, the semantics which the speaker uttered in step S138 is recognized, and the data DJ of the recognition result is further outputted to the signal-processing section 9 in step S140.
[0072] thus, the 1st threshold THD1 -- the logarithm of a high level -- it asks for the section extended by the section of predetermined value taua and taud on the basis of BACKWARD at the FORWARD and termination time the time of being the beginning from which power P (t) was obtained as the last logging section Tc, and the following effect will be acquired, if feature-vector [within this section Tc] V (t) is extracted and speech recognition is carried out.
[0073] the case where the noise of a high level becomes a threshold THD1 from a threshold THD2 in drawing 8 , without becoming low from a threshold THD2 -- this noise -- the logarithm of a voice component -- it judges with power P (t), and carries out, and the feature vector by the noise is memorized in a storage region MEM. however, a time -- FORWARD -- being based -- having asked -- presumption -- initiation -- a time -- criteria [ta] -- carrying out -- the logarithm after it -- since power P (t) is judged

as power of a voice component, the feature vector by the noise can be excepted from the object of speech recognition, and mixing of a noise component can be suppressed to the minimum.
[0074] furthermore, the case where the noise of a high level mixes from a threshold THD2 [near the D] at the time in drawing 8 -- this noise -- the logarithm of a voice component -- it judges as power P (t) and the feature vector by the noise is memorized in a storage region MEM. however, a time -- BACKWARD -- being based -- having asked -- presumption -- termination -- a time -- criteria [td] -- carrying out -- the logarithm before it -- since power P (t) is used as a voice component, the feature vector by the noise can be excepted from the object of speech recognition, and mixing of a noise component can be suppressed to the minimum.
[0075] Thus, according to this operation gestalt, even if it is perimeter environment with many noises, mixing of a noise can be controlled to the minimum and the rate of speech recognition can be raised. The voice detection and the extract which conformed to practical use especially are attained.
[0076]
[Effect of the Invention] According to this invention, the level of the power component signal generated from a sound signal as explained above The 1st threshold, Based on the 2nd threshold of a low, compare from the 1st threshold, and among power component signals, are a high level from the 2nd threshold, and the period which the power component signal which changes to a high level produces is continuously detected from the 1st threshold in time. Since the sound signal in the period is detected as utterance voice and extracted, while being able to suppress mixing of a noise, utterance voice can be detected faithfully and can be extracted.
[0077] Moreover, since the sound signal which exists within limits which carried out predetermined period width-of-face expansion from the first transition portion, the trailing-edge portion or first transition portion, and trailing-edge portion of the power component signal which exceeds the 1st threshold among sound signals was made into the sound signal of final utterance voice, even when a noise is made, the utterance voice of a main part can be detected and extracted in a high precision.
[0078] Moreover, since the feature-extraction means was made to perform an audio feature extraction based on the extracted sound signal synchronizing with the above-mentioned voice section decision means extracting the sound signal of the above-mentioned utterance voice, high-speed speech recognition can be performed.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

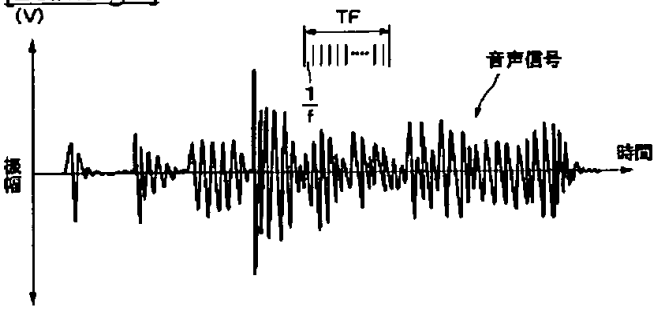
[Brief Description of the Drawings]
[Drawing 1] It is the block diagram showing the configuration of the voice detection equipment concerning this operation gestalt.
[Drawing 2] It is a flow chart for explaining the actuation in the 1st operation gestalt.
[Drawing 3] a logarithm -- it is a wave form chart for explaining the generation method of power.
[Drawing 4] a logarithm -- it is the wave form chart having shown the temporal response of power typically.
[Drawing 5] It is explanatory drawing showing the memory map of the storage section.
[Drawing 6] It is a flow chart for explaining the actuation in the 2nd operation gestalt.
[Drawing 7] It is a flow chart for explaining the actuation in the 3rd operation gestalt.
[Drawing 8] the logarithm in the 3rd operation gestalt -- it is the wave form chart having shown the temporal response of power typically.
[Drawing 9] It is a wave form chart for explaining the trouble of the conventional technology.
[Description of Notations]
1 -- Voice detection equipment
2 -- Microphone
3 -- Front-end processing section
4 -- a logarithm -- power operation part
5 -- Voice section decision section
6 -- Feature-extraction section
7 -- Storage section
8 -- Speech recognition section
RAM -- Memory which can be restored
ROM -- Read-only memory

[Translation done.]

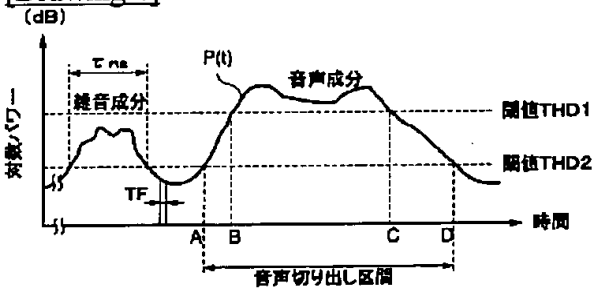
* NOTICES *

DRAWINGS

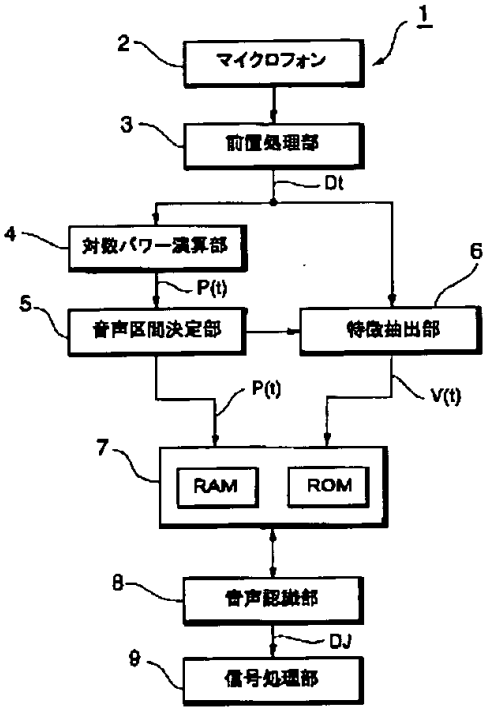
[Drawing 3]



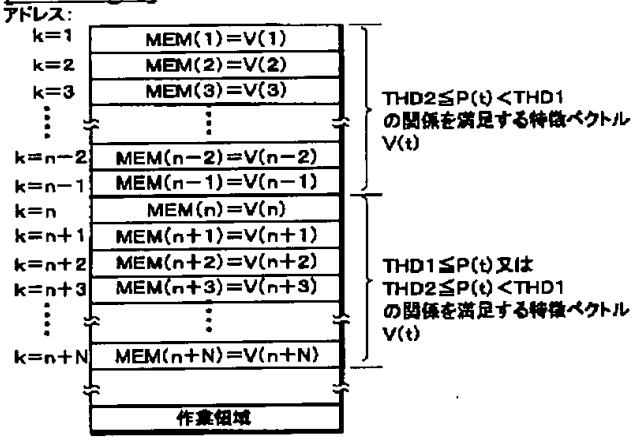
[Drawing 4]



[Drawing 1]

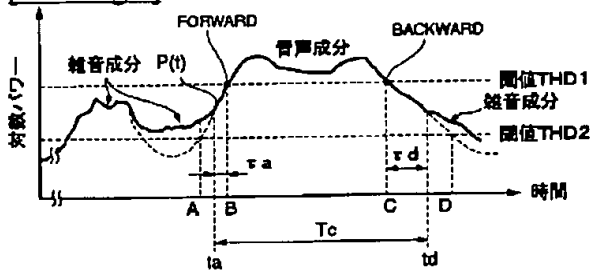


[Drawing 5]

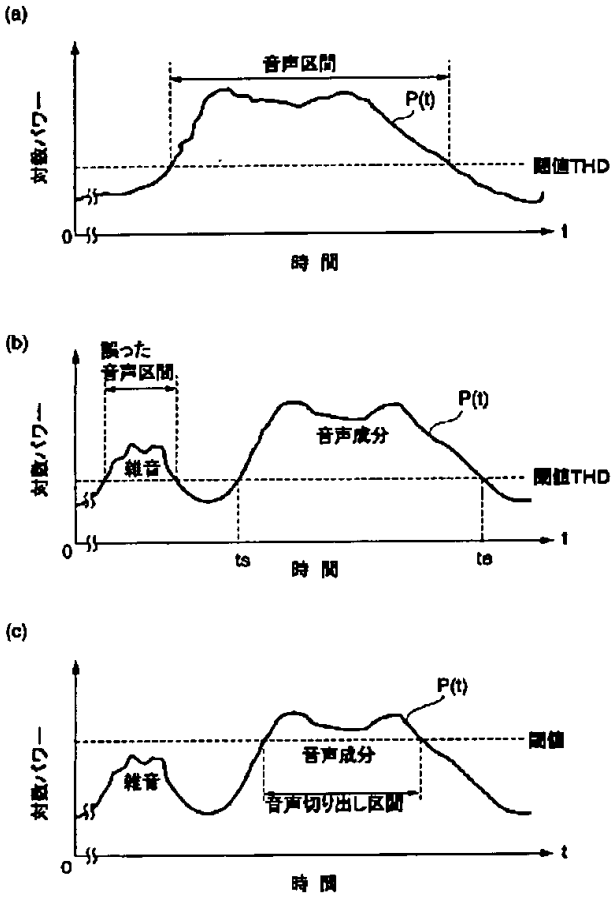


(記憶領域MEMのメモリマップを示す図)

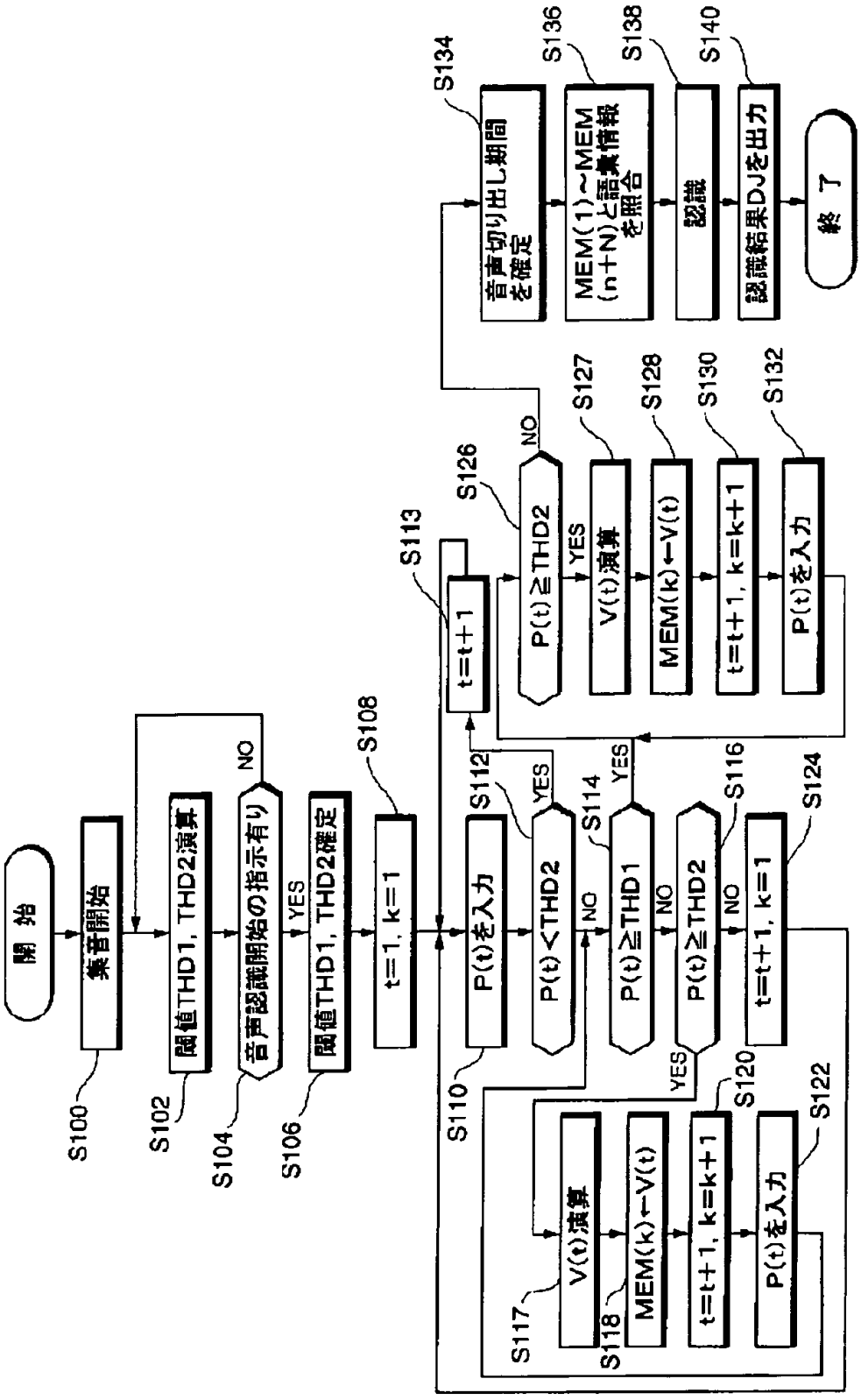
[Drawing 8]



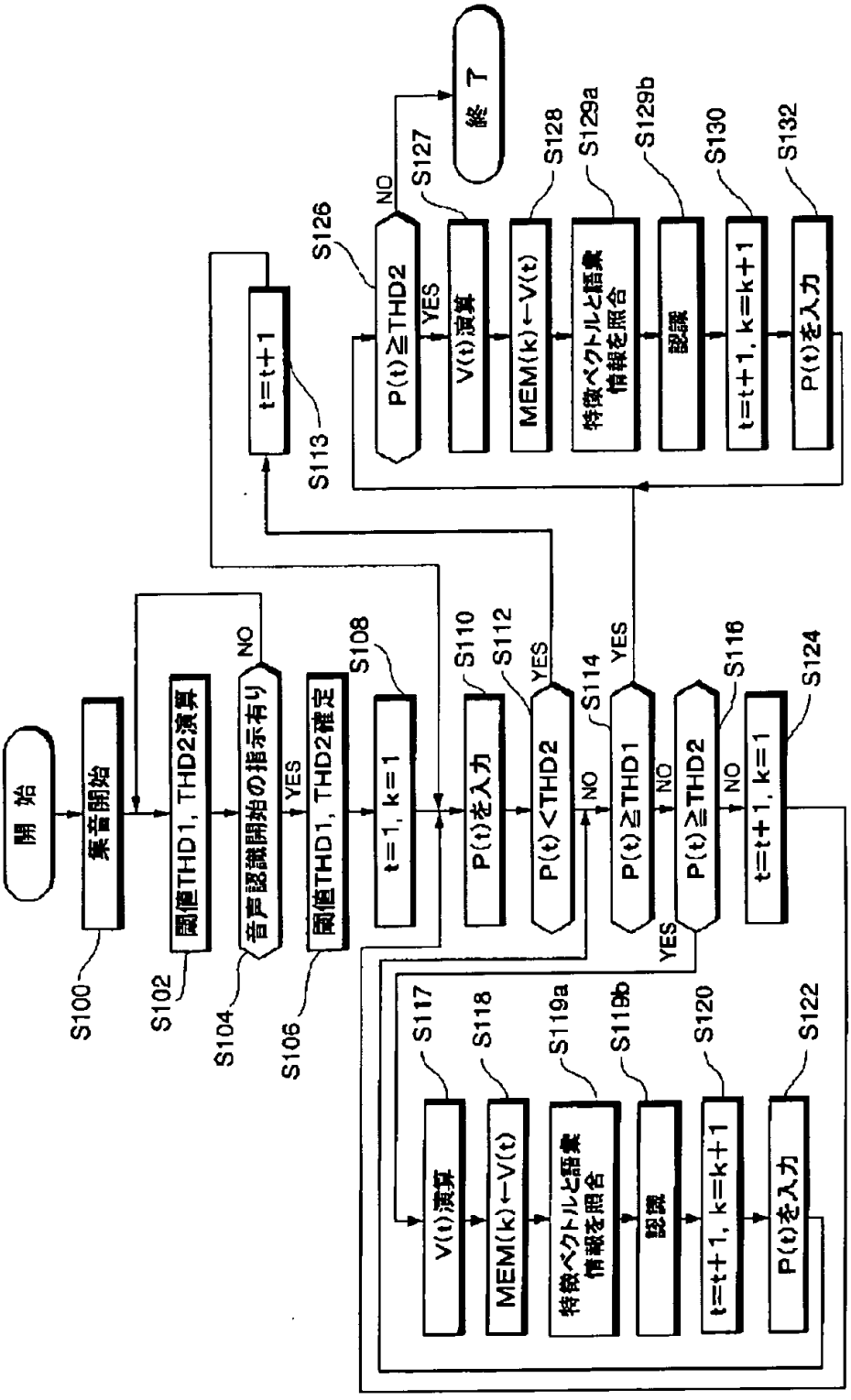
[Drawing 9]


(a)
音声区間
P(t)
閾値THD
対数パワー
0
t
時 間
(b)
誤った
音声区間
P(t)
音声成分
雑音
閾値THD
対数パワー
0
ts
te
t
時 間
(c)
P(t)
閾値
音声成分
雑音
音声切り出し区間
対数パワー
0
t
時 間

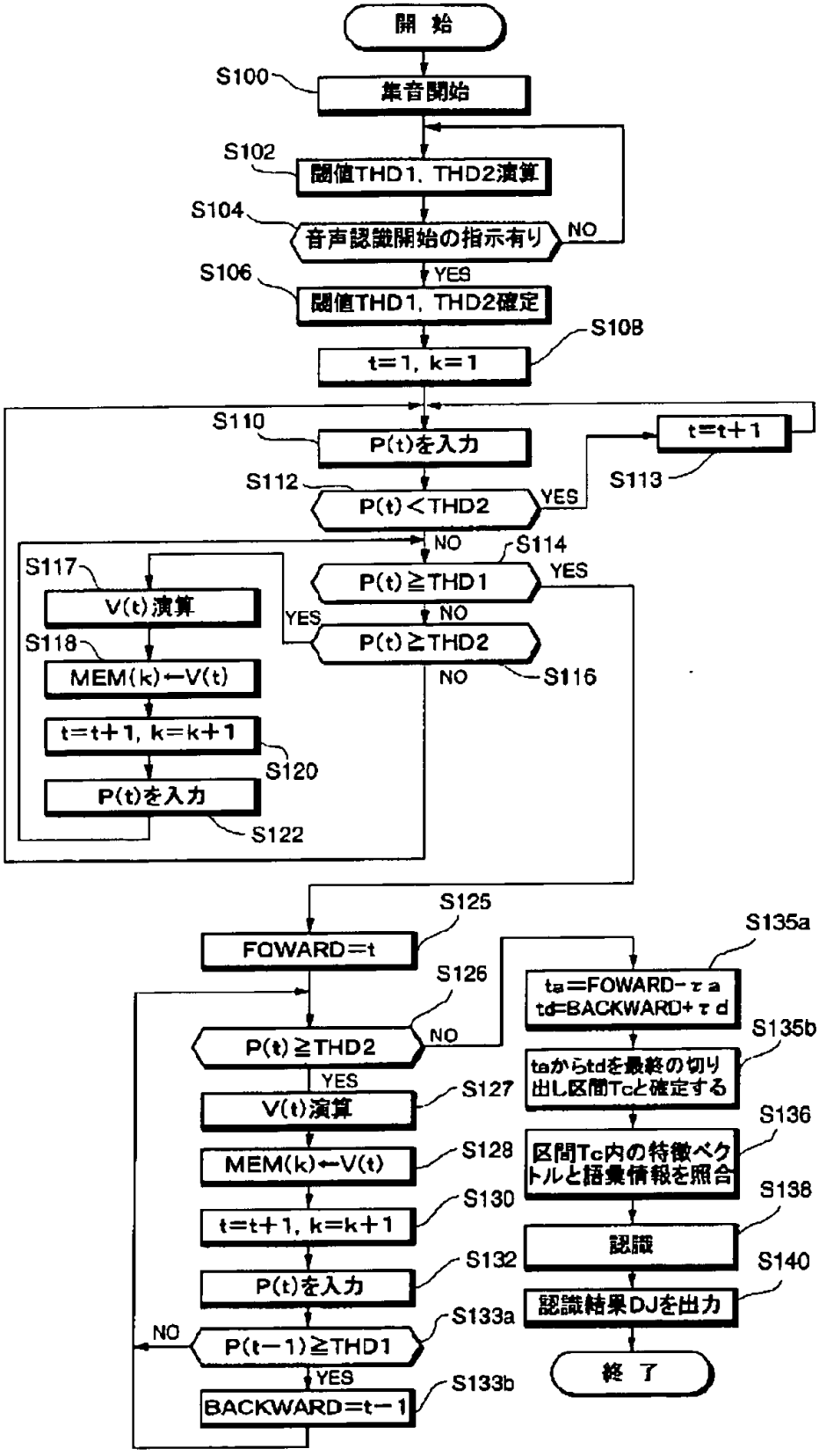
[Drawing 2]


開始
集音開始
S100
閾値THD1, THD2演算
S102
音声認識開始の指示有り
NO
YES
S104
閾値THD1, THD2確定
S106
t=1, k=1
S108
P(t)を入力
S110
P(t)<THD2
YES
S112
t=t+1
S113
NO
P(t)≧THD1
YES
S114
NO
P(t)≧THD2
YES
S116
NO
t=t+1, k=1
S124
P(t)≧THD2
S126
NO
YES
V(t)演算
S127
MEM(k)←V(t)
S128
t=t+1, k=k+1
S130
P(t)を入力
S132
V(t)演算
S117
MEM(k)←V(t)
S118
t=t+1, k=k+1
S120
P(t)を入力
S122
音声切り出し期間を確定
S134
MEM(1)～MEM(n+N)と語彙情報を照合
S136
認識
S138
認識結果DJを出力
S140
終了

[Drawing 6]


開 始
集音開始
閾値THD1, THD2演算
音声認識開始の指示有り
閾値THD1, THD2確定
t=1, k=1
P(t)を入力
P(t)<THD2
P(t)≧THD1
P(t)≧THD2
t=t+1, k=1
V(t)演算
MEM(k)←V(t)
特徴ベクトルと語彙情報を照合
認識
t=t+1, k=k+1
P(t)を入力
P(t)≧THD2
V(t)演算
MEM(k)←V(t)
特徴ベクトルと語彙情報を照合
認識
t=t+1, k=k+1
P(t)を入力
t=t+1
終 了
YES
NO
S100
S102
S104
S106
S108
S110
S112
S114
S116
S117
S118
S119a
S119b
S120
S122
S124
S126
S127
S128
S129a
S129b
S130
S132
S113

[Drawing 7]



[Translation done.]